Modèles de mélanges topologiques pour la classification de données catégorielles et mixtes

Nicoleta Rogovschi*,Mustapha Lebbah**, Younès Bennani**

*LIPADE, Université Paris Descartes
45 rue des Saints Pères
75270 Paris Cedex 06
France
**LIPN-UMR 7030 Université Paris 13 - CNRS
99, av. J-B Clément - F-93430 Villetaneuse France.
prénom.nom@lipn.univ-paris13.fr

Résumé. Cet article présente une méthode basée sur les cartes auto-organisatrices probabilistes dédiées à la classification non supervisée et la visualisation de données catégorielles et des données mixtes contenant des composantes quantitatives et binaires. Pour chacun de ces types de données, nous proposons un formalisme probabiliste dans lequel les unités de la carte topologique sont représentées par un modèle de mélanges de loi de Bernoulli, dans le cas des données binaires et par un modèle de mélanges de lois de Bernoulli et Gaussienne dans le cas des données mixtes. Dans cette étude, la carte topologique est vue comme un modèle génératif et est revisitée dans un formalisme probabiliste de modèles de mélanges. L'idée de base de ce travail repose sur le principe de la conservation de la structure initiale des données en utilisant le formalisme probabiliste. Les modèles de mélanges proposés ici vérifient ce principe et fournissent des résultats directement interprétables par rapport aux données initiales, qu'elles soient simplement binaires ou mixtes. L'apprentissage consiste alors à estimer les paramètres du modèle en maximisant la vraisemblance des données d'apprentissage. L'algorithme d'apprentissage (PrMTM :Probabilistic Mixed Topological Map) que nous proposons est basé sur l'algorithme EM (Estimation-Maximisation). Nous avons montré que l'algorithme à base de modèles de mélanges fournit différentes informations pertinentes qui peuvent être utilisées dans des applications pratiques. Nos approches ont été validées sur différentes bases de données réelles et fournissent des résultats prometteurs.

1 Introduction

L'apprentissage non supervisé consiste à construire des représentations simplifiées de données, pour mettre en évidence les relations existantes entre les caractéristiques relevées sur des données et les ressemblances ou dissemblances de ces dernières, sans avoir aucune connaissance sur les classes. On peut distinguer deux grandes familles : les méthodes probabilistes et les méthodes déterministes ou tout simplement les méthodes de quantification. Ce travail concerne le